A UNIFIED APPROACH TO SCENE CHANGE DETECTION IN UNCOMPRESSED AND COMPRESSED VIDEO

W. A. C. Fernando, C. N. Canagarajah and D. R. Bull
Image Communications Group, Centre for Communications Research
University of Bristol, Merchant Ventures Building
Bristol BS8 1UB, United Kingdom
E-mail: W.A.C.Fernando@bristol.ac.uk

ABSTRACT

There is an increasing need to extract key information automatically from video for the purposes of indexing, fast retrieval and scene analysis. To support this vision, reliable scene change detection algorithms must be developed. This paper describes a unified approach for scene change detection in uncompressed and MPEG-2 compressed video sequences using statistical properties of each image. An efficient algorithm is proposed to estimate statistical features in compressed video without full frame decompression and used these features with the uncompressed domain algorithms to identify scene changes in compressed video. Proposed scheme aims at detecting abrupt transitions and gradual transitions in both uncompressed and MPEG-2 compressed video using a single framework. Results on video of various content types are reported and validated. Furthermore, results show that for uncompressed video the accuracy of the detected transition region is above 98% and above 95% for MPEG-2 compressed video.

1. INTRODUCTION

Due to rapid advances in video products such as digital cameras, camcorders, storage devices (DVDs) and the explosion of the internet, the digital video "for every one" is now becoming a reality. The demand for digital video is also increasing in areas such as video conferencing, multimedia authoring systems, education and video-on-demand systems. Today, the major bottleneck preventing the widespread use of digital video is the ability to find a desired information from a huge database using content. A reliable way to solve this issue is to index the video sequence properly, thus enabling fast access to the video clips stored in multimedia databases.

An important initial task in video indexing is to partition video into relevant temporal segments. The most famous approach of content based video segmentation is shot transition detection in which a video sequence is partitioned into shots, where each video shot represents a meaningful event or a continuous sequence of action. Shot transitions can be divided into two categories: abrupt transitions and gradual transitions. Gradual transitions include camera movements: panning, tilting, zooming and video editing special effects: fade-in, fade-out, dissolving, wiping. Both these transitions are used in narrative film and video to convey story structure. Figure 1 shows the hierarchical description of video and examples for sudden and gradual transitions. Therefore, the ability to identify shot transitions automatically is the first step towards automatic video indexing or video storyboard browsing.

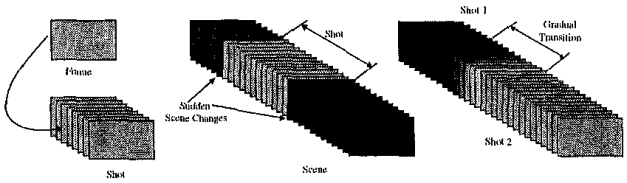


Figure 1: A Hierarchical description of video sequence

The large channel bandwidth and memory requirements for the transmission and storage of image and video necessitate the use of video compression techniques [1]. Hence, the visual data in multimedia databases is expected to be stored mostly in the compressed form. Therefore, finding a desired information from a compressed database is becoming more important. In order to avoid the unnecessary decompression operations in indexing and searching processes, it is efficient to index the image and video in the compressed format. Thus, a powerful scene transition detection algorithm, which can operate both in uncompressed and compressed domain, is required to allow for a complete characterisation of the video sequences. In this paper a hierarchical unified approach is proposed for scene change detection in both uncompressed and compressed video sequences using statistical features and structural properties of each image.

Rest of the paper is organised as follows: In section 2 several related works on scene change detection in uncompressed and compressed video are briefly reviewed. Section 3 describes suitable mathematical models for video special effects. The concept of S-sequences and how to extract them from compressed domain are presented in section 4. Proposed scheme for scene change detection is presented in section 5. Experimental results are given in section 6. Finally, section 7 presents the conclusions and the future work.

2. RELATED WORK

Abrupt transitions are very easy to detect as the two successive frames are completely uncorrelated. Gradual transitions are more difficult to detect as the difference between frames corresponding to two successive shots is substantially reduced. The number of possible gradual transitions is quite high. However, in practice most gradual transitions fall into either fade/dissolve or wipe transitions.

Original manuscript received June 19, 2000

Considerable work has been reported on detecting abrupt transitions both in uncompressed video [2] and compressed video [3,4,5,6,7]. The most famous approaches include histogram difference [2], frame difference, motion vector analysis for uncompressed video [4] and DC-components [6,7], DC-sequences [5] and number of interpolated blocks [3] for compressed video.

Substantial effort has also been devoted toward gradual scene change detection [2,5,8,9,10,11,12] in uncompressed video. For the detection of gradual scene changes twin-comparison method has been proposed [2]. This takes into account the cumulative differences between the frames and requires two cut-off thresholds, one higher threshold (T_h) for detecting abrupt transitions and a lower one (T_l) for gradual transitions. In the first stage a higher threshold is used for detection of abrupt transitions. In the next stage a lower threshold is used and any frame that has the difference more than this threshold is declared as potential start of the transition. However, most of gradual transitions cannot be detected with twin comparison as the difference falls below the lower threshold [2]. Furthermore, this scheme is not suitable for real time processing or to classify gradual transitions. Comparison based on successive frames alone will not be adequate for the detection of gradual transitions because changes are small in this case. One alternative is to use every k^{th} frame instead, i.e., to perform temporal sub-sampling [5]. In this scheme [5], every frame is used and compared to the following k^{th} frame. The main problem with this scheme is to select the value for k. Another limitation of this scheme is that it cannot classify the type of gradual transition.

2.1 Dissolve/fade detection

Zabith et al [11] proposed a feature-based algorithm for detecting and classifying scene beaks. Dissolve and fade are identified by looking at the relative values of the entering and exiting edge percentages. This algorithm requires edge detection in every frame, which is very costly. Another limitation of this scheme is that the edge detection method does not handle rapid changes in overall scene brightness, or scenes with high contrast. Furthermore, automatic segmentation and classification is not possible with this scheme [8]. Alattar used quadratic behaviour of the variance curve to detect fading [8]. This algorithm can only detect fade-in and fade-out where the end frames are fixed. When the sequence has considerable motion, this algorithm fails to identify fade-in and fade-out regions. We have previously considered the ratio between incremental change in the mean of the luminance signal to the same for the chrominance as the criteria of identifying fade transitions [9]. This algorithm may fail to identify fade regions when the solid colour is very close to the mean of the original sequence (before fading is applied).

2.2 Wipe detection

Several algorithms for wipe scene change detection in uncompressed video have also been proposed [10,12]. Alattar proposed statistical feature based approach for wipe detection [10]. This scheme is very sensitive to the type of video sequence as the algorithm is proposed under a crude approximation for the mean and variance curves. Furthermore, the performance of this algorithm is also highly sensitive to motion in the scenes comprising the wipe region. Yu *et. al.* proposed a multi-resolution video segmentation scheme for wipe transition identification [12]. There are two drawbacks in this technique. They assumed that there is no local motion during the wipe transition. In addition, this scheme did not deal with subregion translation and multiregion translation. Furthermore, previous approaches failed to identify the nature of wiping such as wiping pattern and wiping direction.

Above discussion reveals that most algorithms are designed only for specific scene change detection. Another drawback of the existing algorithms is that most of them cannot work in compressed domain. Therefore, an algorithm, which can operate within a single framework for both uncompressed and compressed domain, would be highly desirable for content-based video indexing and retrieval. This paper presents a hierarchical unified approach for scene change detection in both uncompressed and compressed video sequences to resolve these two major problems. In the proposed scheme statistical behaviour and structural properties of each image is used to identify these transitions. Furthermore, an efficient technique is proposed to extract statistical features (mean and variance) from compressed video without full frame decompression, which facilitates the use of the same algorithm used in uncompressed video.

3. MODELING VIDEO SPECIAL EFFECTS

Video special effects are needed to enhance the quality of the video production. Most special effects can be divided into three major categories: dissolving, fading and wiping. All these special effects are used to change the scene gradually between two scenes. These transitions are more difficult to detect as the difference between frames corresponding to two successive shots is substantially reduced. In this section, mathematical models for these special effects are described, which facilitate an analytical solution to the detection problem.

3.1 Dissolving/fading

In video editing and production, proportions of two or more picture signals are simply added together so that the two pictures appear to merge on the output screen. Very often this process is used to move on from picture F to picture G. In this case, the proportions of the two signals are such that as the contribution of picture F changes from 100% to zero, the contribution of picture G changes from zero to 100%. This is called dissolving. When picture F is a solid color, it is called as fade-in and when picture G is a solid colour, it is

known as fade-out. Therefore, dissolving, fade-in and fade-out can be modelled as shown in Equations (1), Equation (2) and Equation (3) respectively.

$$s_n(x,y)=\begin{cases} f_n(x,y) & 0\le n<L_1 \\ \left[1-\left(\frac{n-L_1}{L}\right)\right]f_n(x,y)+\left(\frac{n-L_1}{L}\right)g_n(x,y) & L_1\le n\le(L_1+L) \\ g_n(x,y) & (L_1+L)<n\le L_2 \end{cases} \quad (1)$$

$$s_n(x,y)=\begin{cases} f_n(x,y) & 0\le n<L_1 \\ \left[1-\left(\frac{n-L_1}{L}\right)\right]C+\left(\frac{n-L_1}{L}\right)g_n(x,y) & L_1\le n\le(L_1+L) \\ g_n(x,y) & (L_1+L)<n\le L_2 \end{cases} \quad (2)$$

$$s_n(x,y)=\begin{cases} f_n(x,y) & 0\le n<L_1 \\ \left[1-\left(\frac{n-L_1}{L}\right)\right]f_n(x,y)+\left(\frac{n-L_1}{L}\right)C & L_1\le n\le(L_1+L) \\ g_n(x,y) & (L_1+L)<n\le L_2 \end{cases} \quad (3)$$

where, C is the video signal level (solid colour), $s_n(x,y)$ is the resultant video signal, $f_n(x,y)$ is picture F, $g_n(x,y)$ is picture G, L_1 is length of sequence F, n is the frame number, L is length of dissolving/fading sequence and L_2 is length of the total sequence.

3.2 Wiping

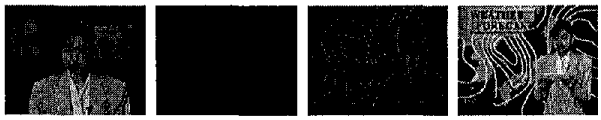
Wiping is a transition from one scene to another where the new scene is revealed by a moving boundary. This moving boundary can be any geometric shape. However in practice this geometric shape is either a line or a set of lines. According to the geometric shape of this boundary, there are about 20-30 different moving boundaries used for wiping in video production. Wiping can be modeled as shown in Equation (4).

$$s_n(x,y)=\begin{cases} f_n(x,y) & n<L_1 \\ P_n\otimes f_n+\overline{P}_n\otimes g_n & L_1\le n\le(L+L_1) \\ g_n(x,y) & (L+L_1)<n\le L_2 \end{cases} \quad (4)$$

where, "$\otimes$" denotes element by element matrix multiplication and matrix P generates the wiping pattern, P_n matrix represent the wiping transition (elements of P_n are either "1" or "0" always).

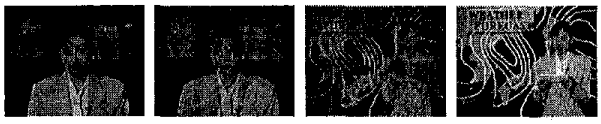
Figure 2 illustrates the behaviour of these shot transitions visually.

Abrupt transition: *A sudden scene change*



Fade-in: *Second scene gradually appearing*

Fade-out: *First scene gradually disappearing*



Dissolve: *First scene gradually disappearing and second scene gradually appearing*

Figure 2: Shot transition illustration

4. STATISTICAL IMAGE AND STATISTICAL-SEQUENCE

A statistical image (S-image) has been defined by spatially reducing the original image into blocks of M'xN' pixels. The (m,n) pixel of the S-image has two elements and defined as in Equations (5) and (6).

$$S'_{\mu,m,n}=\frac{1}{M'N'}\left(\sum_{i=1}^{M'}\sum_{j=1}^{N'}\Omega_{(m-1)M'+i,(n-1)N'+j}\right) \quad (5)$$

$$S'_{\sigma,m,n}=\frac{1}{M'N'}\left(\sum_{i=1}^{M'}\sum_{j=1}^{N'}\left(\Omega_{(m-1)M'+i,(n-1)N'+j}-S'_{\mu,m,n}\right)^2\right) \quad (6)$$

where M',N' - block dimension, (m,n) - Pixel location in the S-image, Ω - original image and S' - S-image.

Sequences formed in such a manner are called as statistical sequences or S-sequences. Thus, generation of S-sequences in uncompressed domain is straightforward. Following sections show how S-sequences are generated for intra-coded and motion compensated frames without full frame decompression. In section 5.3, how these S-sequences are used for wipe scene change detection will be explained.

4.1 Estimation of statistical features in intra-coed images

MPEG-2 compressed video is considered in this paper as it is used in many current and emerging products. Currently, existing image processing operations performed on the compressed sequences require full frame decompression. To maximise the benefits from data compression, it would be advantageous to develop processing algorithms that do not require decompression of the entire compressed data. Operations on compressed bit streams directly or with minimal decoding of relevant information will then eliminate the computational time necessary for full decompression and the extra storage needed for the decompressed results. In this section, an effective technique for extracting mean and variance is proposed for intra-coded images.

In general, a macroblock (MB) of size N_1xN_2 is considered. Let, $s(k,i,j)$ and $S(k,i,j)$ be the uncompressed image and compressed image respectively. Here, k is the MB number ($k=1:N$, N is the total number of MBs in an image) Let, $\mu_{k,s}$ and $\sigma^2_{k,s}$ be mean and the variance of the k^{th} MB in

uncompressed image respectively. Then, mean and variance of the k^{th} MB can be evaluated using Equation (7) and Equation (8) respectively.

$$\mu_{k,s} = \left(\frac{1}{N_1 N_2}\right)\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} s(k,i,j) \tag{7}$$

$$\sigma_{k,s}^2 = \left(\frac{1}{N_1 N_2}\right)\left[\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} s(k,i,j)^2\right] - \mu_{k,s}^2 \tag{8}$$

Using Parseval's theorem, power in uncompressed domain signal can be related to compressed domain signal as shown in Equation (9) assuming that quantisation noise is negligible.

$$\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} s(k,i,j)^2 = \sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} S(k,i,j)^2 \tag{9}$$

It can be proved that mean of the k^{th} MB can be evaluated as follows.

$$\mu_{k,s} = \sqrt{\frac{1}{N_1 N_2}}\, S(k,0,0) \tag{10}$$

Therefore, $\sigma_{k,s}^2$ can be expressed as shown in Equation (11).

$$\sigma_{k,s}^2 = \left(\frac{1}{N_1 N_2}\right)\left[\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} S(k,i,j)^2 - S(k,0,0)^2\right] \tag{11}$$

Thus, equation (11) shows that $\sigma_{k,s}^2$ can be calculated by taking the average power of AC-coefficients in each MB. Therefore, mean and variance can be evaluated in compressed domain using equation (10) and (11) respectively without full frame decompression. It can also be shown that mean (μ_s) and variance (σ_s^2) of the whole image, which contains N number of MBs, can be calculated from Equation (12) and (13) as given below.

$$\mu_s = \frac{1}{N}\sqrt{\frac{1}{N_1 N_2}}\sum_{k=1}^{N} S(k,0,0) \tag{12}$$

$$\sigma_s^2 = \left(\frac{1}{N N_1 N_2}\right)\left[\left[\sum_{k=1}^{N}\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} S(k,i,j)^2\right] - \frac{1}{N}\left[\sum_{k=1}^{N} S(k,0,0)\right]^2\right] \tag{13}$$

4.2 Estimation of statistical features in inter-coed images

MPEG-2 defines three types of pictures: Intra pictures (I-Pictures), Predictive pictures (P-Pictures) and Bi-directional pictures (B-Pictures). The extraction of DCT coefficients from intra-coded frames are trivial. However, the extraction of DCT coefficients from inter-coded frames (P and B) are not trivial since these frames are motion compensated. DCT coefficients in inter-coded frames can be extracted from an intra-coded frame [13]. Therefore, both mean and the variance can be estimated as in intra-coded frames. Thus, S-sequence can be generated for both intra and motion compensated frames from Equation (14) and (15).

$$S'_{\mu,m,n} = \frac{1}{k_1 k_2 \sqrt{N_1 N_2}}\sum_{k=1}^{k_1 k_2} S_{(m-1)M'+i,(n-1)N'+j}(k,0,0) \tag{14}$$

$$S'_{\sigma,m,n} = \frac{1}{k_1 k_2 N_1 N_2}\sum_{k=1}^{k_1 k_2}\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} S_{(m-1)M'+i,(n-1)N'+j}(k,i,j)^2 - S'^{\,2}_{\mu,m,n} \tag{15}$$

where, k_1 and k_2 are positive integers and defined as $M' = k_1 N_1$ and $N' = k_2 N_2$ respectively.

5. SCENE CHANGE DETECTION

In the proposed scheme, statistical and structural properties of images are used to identify scene transitions. These features are used in three steps to identify these scene transitions sequentially. Since abrupt transitions are very common in video sequences, proposed scheme searches for abrupt scene changes first. These transitions are identified by calculating mean square error (MSE) using global mean and variance of each image. If there are no sudden scene changes then algorithm checks for the next common scene transitions: dissolve and fade scene changes. This is achieved by considering the ratio between the second derivative of the variance and the first derivative of the mean and spikes of the second derivative of the variance. If a dissolve or fade region is not detected, then wipe transitions are identified using statistical features together with structural properties. Following sub-sections explain these steps in detail.

5.1 Abrupt scene change detection

For abrupt scene change detection, a MSE approach is suggested based on the assumption of uniform second-order statistics over an image. The frames are compared on the basis of the statistical characteristics of their intensity levels. Let μ'_n, μ'_{n+1} be the mean intensity values for two consecutive frames and σ'_n, σ'_{n+1} be the corresponding standard deviations. Then, Equation (16) represents the formula that calculates the MSE between two consecutive frames. If the MSE exceeds a pre-determined threshold (T_{ubrupt}), an abrupt transition is declared. If it is below the threshold, then algorithm searches for fade and dissolve transitions.

$$MSE_n = \left(\mu'_n - \mu'_{n-1}\right)^2 + \left|\sigma'^2_n - \sigma'^2_{n-1}\right| \tag{16}$$

5.2 Dissolve/fade scene change detection

Let, $m_{s,n}$ be the mean of the resultant video sequence (s_n) and $\sigma_{s,n}^2$ be the variance of the resultant video sequence. m_f, m_g and σ_f^2, σ_g^2 be mean and variance of each starting frame respectively. $e_f^m(n), e_g^m(n)$ are the error terms of the frame number n with respect to the start frame of the two sequences F and G. Thus, $e_f^m(L_1) = e_g^m(L_1) = 0$ and $e_f^\sigma(L_1) = e_g^\sigma(L_1) = 0$. It can be assumed that the two sequences F, G, error signals and the start frames are mutually exclusive. Then, Equation (17) and (18) show the behaviour

of mean and variance of the dissolved sequence. These two equations show that mean and the variance have linear and quadratic behaviour with additive noise during dissolve operation. Equation (19) describes that in the second derivative of the variance curve has large negative spikes at the start and the end of the dissolve operation.

$$m_{s,n} = \begin{cases} m_f + e_f^m(n) & 0 \le n < L_1 \\ (m_f - \rho_{dis} L_1) + \rho_{dis}\, n + e_m(n) & L_1 \le n \le (L_1 + L) \\ m_g + e_g^m(n) & (L_1 + L) < n \le L_2 \end{cases} \tag{17}$$

$$\sigma_{s,n}^2 = \begin{cases} \sigma_f^2 + e_f^\sigma(n) & 0 \le n < L_1 \\ \lambda_{dis} + e_\sigma(n) & L_1 \le n \le (L_1 + L) \\ \sigma_g^2 + e_g^\sigma(n) & (L_1 + L) < n \le L_2 \end{cases} \tag{18}$$

where $\rho_{dis} = \dfrac{(m_g - m_f)}{L}$, $\lambda_{dis} = \xi n^2 - \left(\dfrac{2\sigma_f^2}{L} + 2L_1\xi\right)n + \left(\sigma_f^2 + L_1^2\xi + \dfrac{2L_1\sigma_f^2}{L}\right)$,

$\xi = \left(\dfrac{\sigma_f^2 + \sigma_g^2}{L^2}\right)$, $e_m(n) = e_f^m(n) - \left(\dfrac{e_f^m(n) - e_g^m(n)}{L}\right)(n - L_1)$, $\xi_e = \left(\dfrac{e_f^\sigma(n) + e_g^\sigma(n)}{L^2}\right)$

$e_\sigma(n) = \xi_e n^2 - 2\left(\dfrac{e_f^\sigma(n)}{L} + L_1\xi_e\right)n + \left(e_f^\sigma(n) + L_1^2\xi_e + \dfrac{2L_1 e_f^\sigma(n)}{L}\right)$

$$\frac{d}{dn^2}\left(\sigma_{s,n}^2\right) \approx \begin{cases} 0 & 0 \le n < L_1 \\ \xi - \dfrac{2\sigma_f^2}{L} \approx -\dfrac{2\sigma_f^2}{L} & n = L_1 + 1 \\ 2\xi & (L_1 + 1) < n \le (L_1 + L) \\ \xi - \dfrac{2\sigma_g^2}{L} \approx -\dfrac{2\sigma_g^2}{L} & n = (L_1 + L) \\ 0 & (L_1 + L) < n \le L_2 \end{cases} \tag{19}$$

Similar equations can be derived for fade-in and fade-out as well and given in appendix. Therefore, following properties can be identified during a dissolve/fade transition (Table 1).

Properties	Dissolve	Fade-in	Fade-out
P1 - Mean	linear with additive noise	linear with additive noise	linear with additive noise
P2 - Variance	quadratic with additive noise	quadratic with additive noise	quadratic with additive noise
P3 - $\sigma_{s,n}^2\big\vert_{\text{start}}$	non-zero	zero	non-zero
P4 - $\sigma_{s,n}^2\big\vert_{\text{end}}$	non-zero	non-zero	zero
P5 - $\left[\sigma_{s,n}^2\right]''\Big\vert_{\text{start}}$	large negative spike	large negative spike followed by a large positive spike	large negative spike
P6 - $\left[\sigma_{s,n}^2\right]''\Big\vert_{\text{end}}$	large negative spike	large negative spike	large positive spike followed by a large negative spike

Table 1: Properties of dissolve/fade transitions

In the proposed scheme, these properties are used to identify dissolving and fading. Although, $\sigma_n^2 = 0$ is a very good indicator to identify start/end of a fade-in/fade-out scene change, this condition may not be satisfied for noisy images. In noisy images all pixels may not have the same luminance value at the start/end frame during a fade-in/fade-out operation. However, start frame of fade-in can be identified using the P-5 property where $\left[\sigma_{s,n}^2\right]''$ has a very large negative spike followed by a very large positive spike. This condition is very robust against noisy images. End of fade-out can also be identified in a similar way. Thus, P-5 property is used with two thresholds to identify start/end frame of fade-in/fade-out scene changes. Large negative spike can be used to identify end/start frame of fade-in/fade-out scene change. Negatives spikes can also be used to identify dissolving. Therefore, using the properties P-5 and P-6 fading and dissolving can be identified. Since these spikes can be occur at anywhere in the sequence, an algorithm, which is only based on these spikes produce too many false alarms. This is because, P-5 and P-6 can be satisfied due to a sudden scene change and due to a large local motion. Therefore, a verification method need to be integrated with properties P-5 and P-6 to identify fading and dissolving successfully. To verify these scene transitions one can consider either mean or variance individually. However, experiments show that considering mean and variance individually will not improve the performance. Therefore, in this proposed scheme both these features are combined in order to improve the performance. A three point moving average filter is used to smooth these statistical features as they can be slightly varied due to motion. Thus the ratio between the second derivative of the variance and the first derivative of the mean (Equation 20) is considered to identify these scene transitions. It can be justified both experimentally and mathematically that this ratio is outperforming any scheme considering either mean or variance individually. Therefore, the absolute value of the differentiation of this ratio ($R(n)$) as defined in Equation 20, is checked to identify these scene transitions with a threshold T_{dis}.

$$R(n) = \left| \frac{d}{dn}\left(\frac{\dfrac{d^2}{dn^2}\left(\bar{\sigma}_{s,n}^2\right)}{\dfrac{d}{dn}\left(\bar{m}_{s,n}\right)} \right) \right| \tag{20}$$

where, $\bar{m}$ and $\bar{\sigma}^2$ describes signal m and σ^2 after filtering.

Therefore, the proposed scheme makes use of properties P-1, P-2, P-5 and P-6 to identify fade-in, fade-out and dissolving scene changes. Initially, algorithm searches for a very large negative spike followed by a very large positive spike and this frame is declared as a potential start frame for a fade-in scene change. If it finds these two spikes, then continuos region, which is identified by Equation (20), is searched for four consecutive frames and if it is satisfied then fade-in operation is confirmed. A continuous region is monitored

until it finds a negative spike to confirm the end of fade-in transition.

If two spikes (negative spike followed by a positive spike) are not detected, then it searches for a negative spike followed by a continuos region for four consecutive frames. This continuous region is monitored until a negative spike or positive spike followed by a negative spike is met. If it detects only a negative spike, then end of dissolve scene change is confirmed. Otherwise, end of fade-out scene change is declared.

If non of the spikes are detected, then there will not be any fade or dissolve transitions and algorithm moves to identify wipe transitions.

5.3 Wiping scene change detection

If it is assumed that both end frames, F and G are fixed frames, then wipes can be detected by taking the difference between $S_n(x,y)$ and $S_{n-1}(x,y)$. Therefore, wipe transition region can easily be distinguished using equation (21). This region will change with the frame number (n) according to the wiping pattern.

$$s_n(x,y) - s_{n-1}(x,y) = \begin{cases} 0 & n < L_1 \\ \nu & L_1 \le n \le (L+L_1) \\ 0 & (L+L_1) < n \le L_2 \end{cases} \quad (21)$$

where, $\nu = (P_n \otimes f_n + \overline{P}_n \otimes g_n) - (P_{n-1} \otimes f_{n-1} + \overline{P}_{n-1} \otimes g_{n-1})$

However, this assumption is not true in practice since certain movements are likely with both F and G due to motion. In this case, computing pixel-wise luminance difference is not sufficient to detect. However this problem can be overcome by introducing S-images to identify wiping. Therefore in this proposed scheme, each original frame in the video sequence is mapped into S-sequence as explained in section 4. Then, Mean Squared Error (MSE) is calculated for corresponding pixels of consecutive frames in the S-sequence to ascertain whether there is a significant change in each pixel of the S-image. A threshold (T_{MSE}) is used to determine the pixels which have changed during the two consecutive images. This threshold is adaptive and is defined as the mean value of all MSE pixels in the S-image (i.e. T_{MSE} = mean (MSE of all pixels in the S-image)). Finally, all MSEs are subjected to T_{MSE} to define a binary image, $diff\_W$ as shown in Equation (22) to identify the exact transition region.

$$diff\_W(n,i,j) = \begin{cases} 1 & if\ MSE_{i,j} > T_{MSE} \\ 0 & if\ MSE_{i,j} \le T_{MSE} \end{cases} \quad (22)$$

where, $i = 1 : \frac{M}{M'}$ and $j = 1 : \frac{N}{N'}$

Identifying the transition region (in S-image) is not sufficient to detect wiping automatically. Transition region consists of a single strip or multiple strips and thickness of a strip can be a single line or multiple lines.

It is clear from the above analysis that wiping pattern can be detected by identifying the strips in transition region. The Hough transform is an established technique, which detects a line or a shape by mapping image edge points into a different space called parametric space [14]. The Hough transform for curve detection is very useful when little is known about the location of a boundary, but its shape can be described as a parametric curve (e.g. a straight line). Its main advantages are that it is relatively unaffected by gaps in curves and by noise. Therefore, we can use the Hough transform with $diff\_W$ to identify the transition region. The number of lines to be detected in the parametric space will depend on the block size. If it is small, a large number of lines need to be detected in order to identify the wiping patterns. This is due to many blocks changing during two consecutive frames. If the block size is large, it may be difficult to identify the blocks, which have changed during wipe transitions. To optimise these two scenarios, parameters have been optimised as $k_1 = 2, k_2 = 2, N_1 = 8$ and $N_2 = 8$. Therefore, it can be proved that the thickness of the strips in the statistical image should either be one or two for 176x144 QCIF sequences considered here. In practice wiping patterns are generated using one, two, three or four moving boundaries. Therefore, in this paper wiping patterns, which are generated using four lines (maximum) are considered. Later we show how this can be extended to other complex wipe patterns as well. Thus, there are eight lines to be detected at the maximum in a binary image. This situation arises when four regions are to be detected and the thickness of each region is two lines (eg. box wipe).

Finally, binary image is subjected to Hough transform and eight highest voted candidates in the parametric space are considered to identify these lines in the transition region. Next step is to consider the gradient of the lines and compare against the gradient of the lines detected in the previous frame and wiping is assumed if these gradients are equal. If this condition is satisfied for consecutive four frames, then wiping is declared at frame *n-4*.

Next step is to classify the wiping patterns. For the classification, number of lines and the combination of gradient and constant values are used. This step is only necessary if wiping is identified. Direction can be identified using the behaviour of the constants. Constant of a detected line is used for this purpose. Therefore the direction is determined as follows.

Forward/outward	Constant is increasing against frame number
Backward/inward	Constant is decreasing against frame number

Algorithm can be extended to identify complex wiping patterns within the same framework of the proposed algorithm with slight modification to the region detection.

For instance, circular and elliptical wipe patterns can be handled by extending the parametric space (Hough transform) to the third dimension. Pentagon and star wipe can be identified by increasing the number of lines to be detected in parametric space. However, these patterns are not very common in video production.

5.4 Scene change detection in compressed domain

In compressed domain statistical features are extracted as explained in section 4 and algorithms described in section 5.1-5.3 are applied to identify scene transitions. Figure 3 explains the complete algorithm for scene change detection in uncompressed and compressed video sequences.

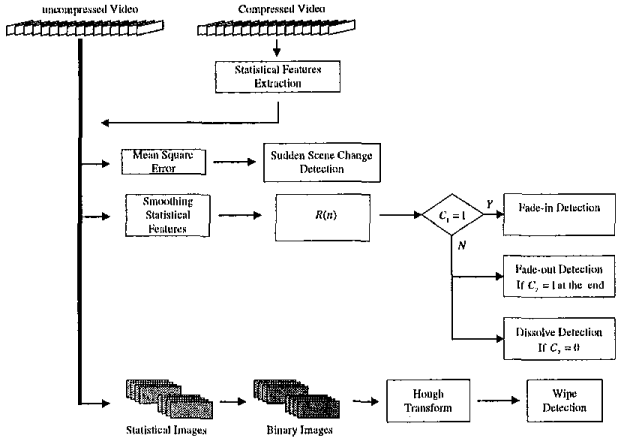


Figure 3: Complete algorithm

Where, $C_1 = 1$ if second derivative of variance find a large negative spike followed by a large positive spike, else $C_1 = 0$. $C_2 = 1$ if second derivative of variance find a large positive spike followed by a large negative spike, else $C_2 = 0$.

6. RESULTS

Proposed algorithm is validated by experiments with a variety QCIF format video sequences.

6.1 Performance of the compressed domain statistical feature extraction model

In this section, results are presented with the proposed feature extraction scheme for compressed data. Several test sequences are considered and compared the estimated variance was compared to the actual variance. Figure 4 shows the estimated standard deviation for MPEG-2 compressed format against the actual standard deviation for ***akiyo*** QCIF test sequence. Extracted standard deviation is very close to the actual curve. Therefore, these estimated features can be used effectively for scene transition detections in compressed domain.

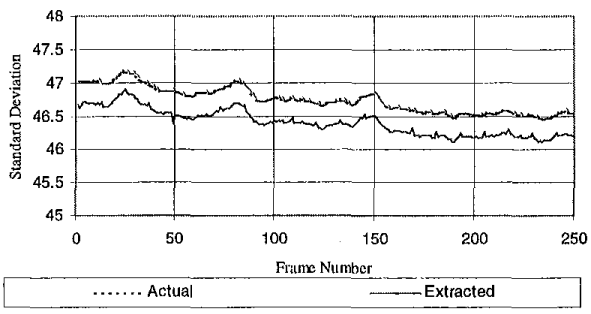


Figure 4: Estimated variance for ***akiyo*** sequence (sub-GOP length = 3, GOP length = 12)

6.2 Sudden scene change detection

The methodology described in section 5.1 has been applied to a test video sequence. Figure 5 shows the MSE variation for the first test sequence. A predetermined threshold (T_{abrupt} =1000) is used to identify abrupt transitions. Summarised results are presented in section 6.5.

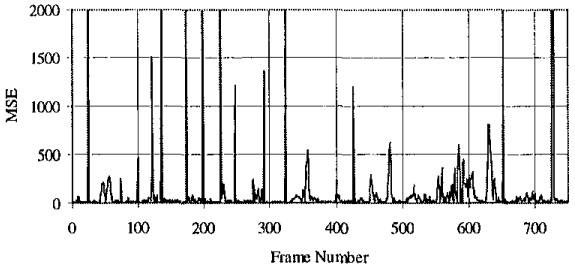


Figure 5: MSE variation

6.3 Dissolve/fade scene change detection

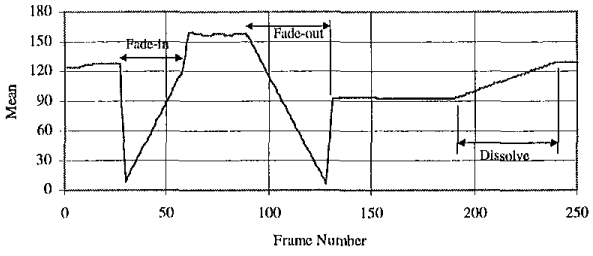


Figure 6: Mean of the luminance signal

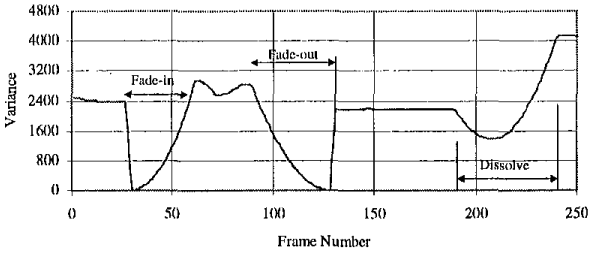


Figure 7: Variance of the luminance signal

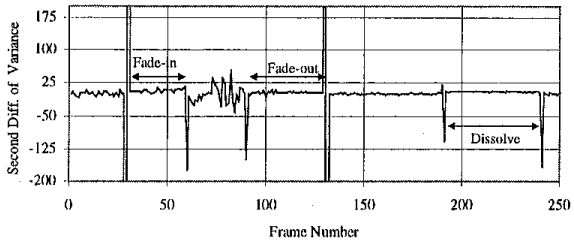


Figure 8: Second derivative of variance of the luminance signal

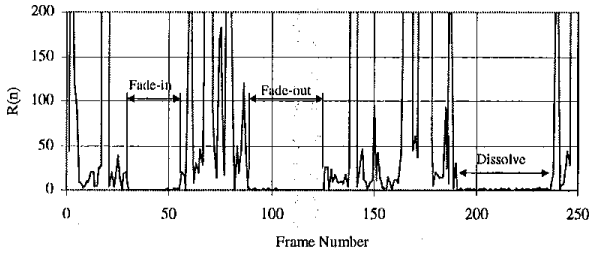


Figure 9: $R(n)$ of the signal

Actual fade region	Detected region (U)	Detected region (C)	Nature of the region
31-60	31-60	31-60	Fade-in
91-130	91-130	91-130	Fade-out
191-240	191-240	191-240	Dissolve
271-330	271-330	271-330	Dissolve
391-440	391-440	391-440	Fade-out
501-580	501-580	501-580	Fade-out
801-836	801-836	801-836	Fade-in
860-910	860-910	860-910	Fade-in
945-986	945-986	945-986	Fade-out
1021-1092	1021-1092	1021-1092	Dissolve
1100-1160	1100-1160	1100-1160	Fade-in
1500-1548	1500-1548	1500-1548	Fade-in
1600-1632	1600-1632	1600-1632	Dissolve
1724-1764	1724-1764	1724-1764	Fade-in
1860-1894	1860-1894	1860-1894	Fade-out

Table 2: Dissolve/fade region identification (U-uncompressed video, C-compressed video)

Proposed algorithm for dissolve and fade scene change detection is validated by experiments with a variety of QCIF format video. Figure 6, 7, 8 and 9 show the mean, variance, second derivative of variance and $R(n)$ respectively for the initial part of the second test video sequence. Figure 8 clearly shows the negative and positive spikes at the start and end of these transitions. From Figure 9, three regions are identified with the threshold T_{dis}=10. Considering the variance of the sequence fade-in, fade-out and dissolve regions are identified correctly. Therefore fade-in, fade-out and dissolve regions are identified from 31st frame to 60th frame, from 91st frame to 130th frame and from 191st frame to 240th frame respectively.

Table 2 summarises some results with the proposed algorithm for dissolve/fade transitions. It indicates that fade and dissolve regions can be detected very accurately both in uncompressed and compressed domain.

6.4 Wipe scene change detection

Consider an example to describe the performance of the algorithm. Figure 10 presents the MSE variation for S-image in frame *n-1* and frame *n*. Figure 10 also describe how the MSE variation is mapped to a binary image with an adaptive threshold. Finally this binary image is subjected to Hough transform to evaluate the average gradient and the constant. Figure 10 also shows how average gradient and the constant is varying for frame *n-1* and frame *n*. It shows that the average gradient is constant (180^0) and the constant is increasing.

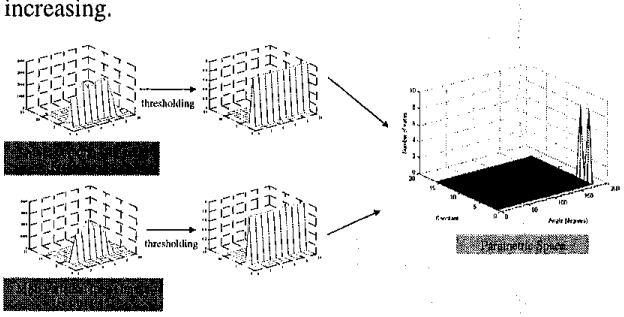


Figure 10: MSE distribution and parametric space

A test sequence is considered here to evaluate the performance of the proposed scheme. Figure 11 and Figure 12 show how the average gradient and the constant of the lines vary with the frame number for the initial part of the third test sequence. From Figure 11 wiping patterns are identified as vertical and horizontal-barn-doors wiping since the average gradient is 180^0 and 90^0 respectively. Since constants are increasing and decreasing (during the period of wiping) with the frame number, wiping directions are identified as forward and in-word respectively. Therefore, forward vertical wiping pattern from 41st frame to 86th frame and in-word horizontal-barn-doors wiping from 100th frame to 141st frame are identified respectively. Even if little is known about the location of a boundary, its shape can be described as a parametric curve since Hough transform is used for line detection. If there are two consecutive wipe regions separated by very small gap, they are bridged to form a longer wipe region.

Table 3 shows that the algorithm is capable of detecting all wipe regions accurately even when the video sequence contains other special effects or camera effects. There are some offsets with box and corn wipe detection. This is due to all lines merging at the start and the end of box-(out-word) and box-(in-word) wiping respectively. Similar behaviour can also be expected in corn wiping.

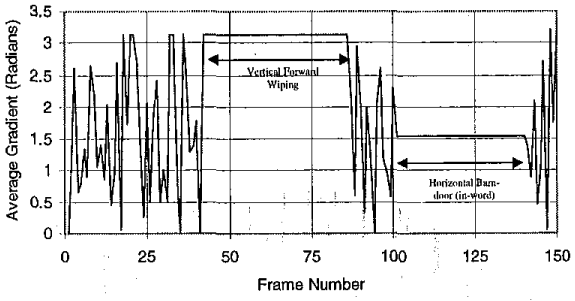


Figure 11: Average gradient of highest voted candidate(s) in parametric space

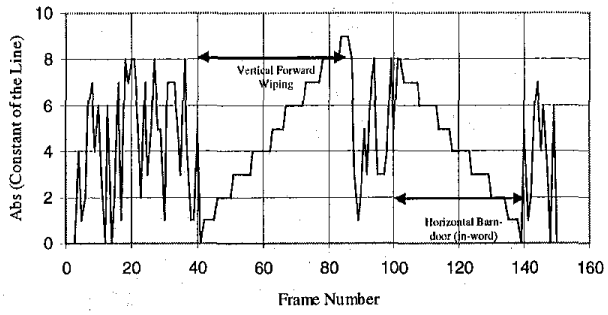


Figure 12: Constant of the highest voted candidate(s) in parametric space

Actual wipe region	Detected wipe region (U)	Detected wipe region (C)	Classification
41-86	41-86	41-86	Vertical-Forward
100-141	100-141	100-141	Horizontal-Barn-doors (in-word)
203-237	203-237	204-237	Horizontal-Backward
340-378	341-378	344-378	Corn-Backward
397-442	397-442	397-441	Horizontal-Barn-doors (In-word)
539-574	539-574	540-574	Horizontal-Barn-doors (out-word)
623-637	623-637	623-636	Horizontal-Barn-doors (in-word)
688-725	688-725	688-724	Horizontal-Forward
868-728	868-726	868-725	Box in-word wiping

Table 3: Wipe region identification

6.5 Summarised results

Table 4 summarises some results with the complete proposed algorithm. Results show that the algorithm is capable of detecting all scene changes accurately. Finally, results are presented in a confusion matrix as shown in table 5. It presents hit rate and the accuracy of the region for all scene changes considered. Accuracy of the region is defined as in the Equation (22). Most scene transitions are identified correctly. Last column shows that two wipe transitions have been missed. These two wipe transitions are circuiar wipe and elliptical wipe transitions. Since the proposed algorithm is only considered most common wipe transitions, it is unable to detect circular and elliptical wipe transitions. Therefore, the proposed algorithm can be used in both uncompressed and compressed video to detect scene changes with a high reliability and less complexity.

Actual transition region	Detected transition region (U)	Detected transition region (C)	Classification
12	12	12	Sudden scene change
45-80	45-80	45-80	Fade-in
240-297	240-297	240-297	Fade-in
369	369	369	Sudden scene change
426-464	426-464	426-464	Horizontal-barn-doors (out-word) wiping
521	521	521	Sudden scene change
592	Missed	Missed	Sudden scene change
680-750	680-750	680-750	Dissolve
824-865	824-865	824-865	Fade-out
980	980	980	Sudden scene change
1078	1078	1078	Sudden scene change
1144-1184	1144-1184	1144-1184	Fade-out
1481	1481	1481	Sudden scene change
1600-1650	1600-1650	1600-1650	Dissolve
1780-1850	1780-1850	1780-1850	Horizontal-backward wiping
2010-2075	2010-2073	2010-2072	Box in-word wiping
2167-2240	2167-2240	2167-2241	Horizontal-forward wiping
2401	2401	2401	Sudden scene change

Table 4: Summarised results

$$\text{Accuracy} = 1 - \frac{\text{Abs(Actual Region} - \text{Detected Region)}}{\text{Actual Region}} \qquad (22)$$

		Abrupt	Fade-in	Fade-out	Dissolve	Wipe
Abrupt	U	24/27, 1.0	-	-	-	-
	C	24/27, 1.0	-	-	-	-
Fade-in	U	-	9/9, 1.0	-	-	-
	C	-	9/9, 1.0	-	-	-
Fade-out	U	-	-	6/6, 1.0		-
	C	-	-	6/6, 1.0	-	-
Dissolve	U	-	-	-	6/6, 1.0	-
	C	-	-	-	6/6, 1.0	-
Wipe	U	-	-	-	-	15/17, 0.98
	C	-	-	-	--	15/17, 0.95

Table 5: Confusion matrix

7. CONCLUSIONS

This paper presented a hierarchical unified approach for scene change detection in both uncompressed and compressed video sequences using statistical features and structural properties of each image. Statistical and structural features, which capture the different characteristics of different kinds of shot transitions, are proposed based on analysis of the production aspects of the video. Section 4 described an efficient technique for extracting variance from MPEG-2 compressed video directly without full frame decompression and used this together with mean for detecting scene transitions in compressed video.

Results show that the algorithm is capable of detecting all scene changes satisfactorily. Therefore, the proposed algorithm can be used in uncompressed and compressed video to detect both sudden and gradual scene changes with a high reliability. Main advantage of the proposed scheme is that they can identify scene transitions within a single framework with less complexity. However, future work is required to extend the algorithm for camera movements detection within the same framework.

APPENDIX

This appendix presents the statistical behaviour for fade-in and fade-out scene changes. Mean and variance of the fade-in sequence can be given by Equation (A-1) and (A-2) respectively. Second derivative of variance is presented by Equations (A-3). Equations (A-4) to (A-6) present the same features for fade-out scene transition.

$$m_{s,n}=\begin{cases} m_f+e_f^m(n) & 0\le n<L_1 \\ (m_f-\rho_{fi}L_1)+\rho_{fi}\,n+e_m(n) & L_1\le n\le(L_1+L) \\ m_g+e_g^m(n) & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-1)}$$

where $\rho_{fi}=\dfrac{(m_g-C)}{L}$ and $e_m(n)=\left(\dfrac{e_g^m(n)}{L}\right)(n-L_1)$

$$\sigma_{s,n}^2=\begin{cases} \sigma_f^2+e_f^a(n) & 0\le n<L_1 \\ \lambda_{dis}+e_a(n) & L_1\le n\le(L_1+L) \\ \sigma_g^2+e_g^a(n) & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-2)}$$

where, $\lambda_{fi}=\xi n^2-(2L_1\xi)n+L_1^2\xi$, $\xi=\left(\dfrac{\sigma_g^2}{L^2}\right)$ $e_a(n)=\xi_e n^2-(2L_1\xi_e)n+L_1^2\xi_e$,

$\xi_e=\left(\dfrac{e_g^a(n)}{L^2}\right)$

$$\frac{d}{dn^2}(\sigma_{s,n}^2)=[\sigma_{s,n}^2]''\approx\begin{cases} 0 & 0\le n<L_1 \\ -\sigma_f^2 & n=L_1 \\ \sigma_f^2 & n=L_1+1 \\ \dfrac{2\sigma_g^2}{L^2} & L_1<n<(L_1+L) \\ -\dfrac{2\sigma_g^2}{L} & n=(L_1+L+1) \\ 0 & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-3)}$$

$$m_{s,n}=\begin{cases} m_f+e_f^m(n) & 0\le n<L_1 \\ (m_f-\rho_{fo}L_1)+\rho_{fo}\,n+e_m(n) & L_1\le n\le(L_1+L) \\ m_g+e_g^m(n) & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-4)}$$

where $\rho_{fo}=\dfrac{(C-m_f)}{L}$ and $e_m(n)=e_f^m(n)-\left(\dfrac{e_f^m(n)}{L}\right)(n-L_1)$

$$\sigma_{s,n}^2=\begin{cases} \sigma_f^2+e_f^a(n) & 0\le n<L_1 \\ \lambda_{dis}+e_a(n) & L_1\le n\le(L_1+L) \\ \sigma_g^2+e_g^a(n) & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-5)}$$

where, $\lambda_{fo}=\xi n^2-\left(\dfrac{2\sigma_f^2}{L}+2L_1\xi\right)n+\left(\sigma_f^2+L_1^2\xi+\dfrac{2L_1\sigma_f^2}{L}\right)$, $\xi=\left(\dfrac{\sigma_f^2}{L^2}\right)$ and

$e_a(n)=\xi_e n^2-2\left(\dfrac{e_f^a(n)}{L}+L_1\xi_e\right)n+\left(e_f^a(n)+L_1^2\xi_e+\dfrac{2L_1e_f^a(n)}{L}\right)$, $\xi_e=\left(\dfrac{e_f^a(n)}{L^2}\right)$

$$\frac{d}{dn^2}(\sigma_{s,n}^2)=[\sigma_{s,n}^2]''\approx\begin{cases} 0 & 0\le n<L_1 \\ -\dfrac{2\sigma_f^2}{L} & n=L_1 \\ \dfrac{2\sigma_f^2}{L^2} & L_1<n<(L_1+L) \\ \sigma_g^2 & n=(L_1+L+1) \\ -\sigma_g^2 & n=(L_1+L+2) \\ 0 & (L_1+L)<n\le L_2 \end{cases} \quad \text{(A-6)}$$

REFERENCES

1. Sikora T., *"MPEG Digital Video-Coding Standards"*, IEEE Signal Processing magazine, pp.82-100, September 1997.
2. Zhang, H.J., *" Automatic Partitioning of Full-Motion Video"*, ACM/Spinger Multimedia Systems, Vol.1, No.1, pp. 10-28, 1993.
3. Fernando, W.A.C., Canagarajah, C.N., Bull, D. R., *"Sudden Scene Change Detection in MPEG-2 Video Sequences"*, Proceedings – IEEE International Workshop on Multimedia Signal Processing, pp. 259-264, 1999.
4. Fernando, W.A.C., Canagarajah, C.N., Bull, D. R., *"Video Segmentation and Classification for Content Based Storage and Retrieval Using Motion Vectors"*, Storage and Retrieval for Image and Video Databases VII, SPIE, pp. 687-698, 1999.
5. Yeo, B.L., Liu, B.,*"Rapid Scene Analysis on Compressed Video"*, IEEE Transactions on Circuits and Systems for video technology, Vol. 5, No 6, pp. 533-544, December 1995.
6. Zhang, H.J., Gong, L.Y., Smoliar, S.W., *" Video Parsing Using Compressed Data"*, Proceedings of IS&T/SPIE, Image and Video Processing II, pp.142-149, February-1994.
7. Meng, J., Juan, Y., Chang, S.F., *" Scene Change Detection in a MPEG Compressed Video Sequence"*, Proceedings of IS&T/SPIE, Vol. 2419, February 1995.
8. Alattar, A.M.,*"Detecting Fade Regions in Uncompressed Video Sequences"*, Proceedings of ICASSP, pp. 3025-3028, 1997.
9. Fernando, W.A.C., Canagarajah, C.N., Bull, D.R., *"Automatic Detection of Fade-in and Fade-out in Video*

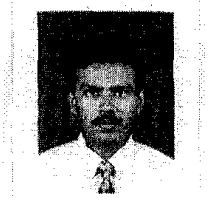
Sequences", Proceedings of ISCAS, Volume IV-Image and Video Processing, Multimedia, and Communications, pp.255-258, 1999.

10. Alattar, A. M., *"Wipe Scene Change Detector For Segmenting Uncompressed Video Sequences"*, Proceedings of ISCAS, pp. 249-252, 1998.
11. Zabith, R., Miller, J., and Mai, K., *"Feature-Based Algorithms for Detecting and Classifying Scene Breaks"*, ACM International Conference on Multimedia, pp. 189-200, California, USA, 1993.
12. Yu, H., Wolf, W. *"A Multi-resolution Video Segmentation Scheme for Wipe Transition Identification"*, Proceedings of ICASSP, pp. 2965-2968, 1998.
13. Chang, S.F., Messerschmitt, D.G., *" Manipulation and composting of MC-DCT compressed video"*, IEEE Journal of Selected Areas in Communications, vol.13, pp.1-11, January 1995.
14. Ballard, D.H. and Brown, C.B., *Computer Vision"*, Prentice-Hall, 1982.

ACKNOWLEDGEMENTS

First author would like to express his gratitude and sincere appreciation to the University of Bristol and CVCP for providing financial support for this work.

BIOGRAPHIES



Anil Fernando received the B.Sc. Engineering degree (First class) in Electronic and Telecommunications Engineering from University of Moratuwa, Sri Lanka in 1995 and the MEng degree (Distinction) in Telecommunications from Asian Institute of Technology, Bangkok, Thailand in 1997. Since 1998, he has been a Ph.D. student at Department of Electrical and Electronic Engineering, University of Bristol. His current research interests include scene change detection in uncompressed and compressed video, video editing in compressed video, intelligent video encoding, COFDM for wireless channels, channel coding and modulation schemes for satellite channels.

Nishan Canagarajah is currently a Senior Lecturer in Signal Processing at University of Bristol. He has BA (Hons) and Ph.D. in DSP techniques for speech enhancement, both from the University of Cambridge. He is a committee member of the IEE Professional Group E5, member of the virtual centre of excellence in digital broadcasting and multimedia technology and an associate editor of the IEE Electronics and Communication Journal. He is also an editor of a book on Mobile Multimedia Technology. His research interests include image and video coding, non-linear filtering techniques and the application of signal processing to audio and medical electronics.

David Bull is currently a Professor of Digital Signal Processing and a University of Bristol research Fellow. He leads the Image Communications Group at Bristol and is Deputy Director of the Centre for Communications Research. He has worked widely in the fields of 1 and 2-D signal processing and his current research is focused on the problems of image and video communications for both low bit rate and broadcast applications. In particular he is working on error resilient source coding, linear and non-linear filter banks, scalable coding methods, motion estimation and architectural optimisation (for filters, transforms and wavelet filter banks).